様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １９ 日現在

機関番号：６２６１８
研究種目：基盤研究(B)（一般）
研究期間：2015～2018
課題番号：１５Ｈ０３２１２

研究課題名（和文）会話文への発話者情報の付与によるコーパスの拡張

研究課題名（英文）Expansion of corpus by annotating speaker's information to conversation sentences

研究代表者
山崎　誠（YAMAZAKI, MAKOTO）

大学共同利用機関法人人間文化研究機構国立国語研究所・言語変化研究領域・教授

研究者番号：３０１８２４８９

交付決定額（研究期間全体）：（直接経費）　　12,800,000円

研究成果の概要（和文）：本研究では，『現代日本語書き言葉均衡コーパス』（BCCWJ）の図書館サブコーパスに含まれる小説（日本文学，英米文学）に話者情報（話者名，性別，年代等）の付与し，それを利用した研究を行った。話者情報は今後，ウェブ上の検索ツール「中納言」に搭載し，検索結果に表示できるようにする。話者情報を利用した研究としては，他の話し言葉のデータ（『日本語話し言葉コーパス』『名大会話コーパス』，『日常会話コーパス』）との比較によって，実際の話し言葉と小説の会話文との違いを語彙的に明らかにした。

研究成果の学術的意義や社会的意義
本研究の学術的意義は，これまで話者情報が付与されていなかった『現代日本語書き言葉均衡コーパス』の小説に対して，話者情報を付与したことである。これにより，実際の話し言葉と会話文のような擬似的な話し言葉の比較が可能になった。本研究の社会的意義として，現在の日本人が持っている，話者とその言葉遣いとの社会的な関係を明らかにできることである。これは役割語研究の深化にも直結するものである。

研究成果の概要（英文）：In this research, we annotated the speaker information (name,gender, age, etc.) to the novels (Japanese literature, English and American literature) contained in the library sub-corpus of "Balanced Corpus of Contemporary Written Japanese"(BCCWJ), and conducted the study using the speaker information. The speaker information will be included in the search tool "Chunagon" on the web near future so that it can be displayed in the search results. As a study using speaker information, comparison between the data of other spoken corpora ("Corpus of Spontaneous Japanese", "Nagoya University Conversation Corpus", "Corpus of Everyday Japanese Conversation Corpus") we clarified the difference lexically.

研究分野：日本語学

キーワード：会話文　小説　話者情報　コーパス　役割語　擬似的話し言葉

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年，日本でも本格的に均衡コーパスを利用した言語研究が盛んになりレジスターやジャンルによる言語使用の違いが解明されつつある。書き言葉においては、応募者（山崎，柏野）がその一員として構築した『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）を利用したレジスター研究が浸透してきている。BCCWJ の言語量の約 6 割を占める書籍のサンプルには，大量の会話文が存在する。会話文は地の文とは言語的に異なる特徴を持つことが多いため，分析に当たっては別に扱うことがあるが，現在の検索環境では地の文と会話文とを区別することが難しい。さらに，会話文には発話者に関する情報が付与されていないため，どのような人物がどういう状況で発した文なのかが分からない不便さがある。BCCWJ を利用した会話文のレジスター的研究は皆無である。一方、British National Corpus (BNC)では、その 1 割（1000 万語）を占める話し言葉部分について<person>タグを付与し，年代，方言，第 1 言語，最終学歴，社会階層，性別などの情報を付与し研究に供している。しかし，書き言葉における会話部分についてはそのようなタグは付与されていない。そこで，本研究では書き言葉における会話文への者情報の付与を試みた。

２．研究の目的

(1)BCCWJ の拡張：

BCCWJ における会話文に対して，話者の属性情報（性別，年代，職業）を付与する。話者の属性は，既存の話し言葉コーパスとの比較も考慮したものにする。

(2)実際の話し言葉と擬似的話し言葉の違いの解明

既存の話し言葉コーパス（『日本語話し言葉コーパス』『名大会話コーパス』）及び構築途中の「日常会話コーパス」と会話文とを比較し，実際の話し言葉と擬似的話し言葉の違いを語彙的・文法的な面から明らかにする。

(3)話者情報を利用した研究

本研究で付与した発話者情報を利用した話者の属性を利用した研究を行う。例えば，男性と女性ではどちらが多く話しているか，あるいは，もともと日本語で書かれた小説と翻訳小説とでは，会話文に違いがあるか，などである。

３．研究の方法

(1)本研究で使用する話者情報（性別，年代，職業等）の設計を行う。

(2)BCCWJ の会話文を抽出し，話者情報を付与する。最初は試行を行い，全体の作業量を試算し，適宜計画を修正する（実際には，試行の結果，図書館サブコーパスの小説に対して話者情報を付与することになった）。

(3)会話文に付与した話者情報をまとめて分析用のデータを作成する。

(4)話者情報を利用し，実際の話し言葉と会話文のような擬似的な話し言葉との違いを語彙的・文法的に解明する。

４．研究成果

(1)話者情報の付与

BCCWJ の図書館サブコーパスに含まれる小説のサンプルは，NDC(図書分類）の 9x3 で表されるが，全体で 2668 サンプルである。そのうち，2659 サンプルに対して話者情報を付与した。話者情報は，話者名，性別（男，女，不明），年代（若年（20 歳未満），成年（20 歳～59 歳），老年（60 歳以上））であるが，この他に多くのサンプルに対して，職業（身分）や会話相手の情報，会話関連情報（方言，電話，引用，外国語，独り言，心内発話）も付与している。

(2)話者情報データの公開

上記の話者情報（話者名，性別，年代）は，2019 年度中にウェブ上の検索ツール「中納言」の検索結果に表示されるようにする。また，BCCWJ 購入者限定であるが，「中納言」のサイトで，上記すべての話者情報データにアクセスすることができる。

(3)話者情報を利用した研究

山崎(2018)（雑誌論文(1)）では，翻訳小説と日本語小説の会話文の違いについて考察した。その結果，以下のことが分かった。①会話文の平均長は日本語小説の方が長い（日本語小説約 978 語，翻訳小説約 709 語）。②品詞構成比では，異なり語数において，日本語小説のほうが名詞が多く，動詞が少ない。③対数尤度比（ＬＬＲ）による特徴語を分析すると，日本語小説には「さん」「さま」「先生」「ちゃん」「君（くん）」などの人につく接尾辞や名詞が有意に多かったのに対し，翻訳小説では，「彼」「私」「君（きみ）」「あなた」などの人称代名詞や「イエス」「イギリス」」「マイケル」などの固有名詞が有意に多かった。

また，山崎（2017）（学会発表（14））では，会話文を他の話し言葉データと比較した。比較対象は『日本語話し言葉コーパス』の学会講演，摸擬講演，『名大会話コーパス』『男性の言葉・女性の言（職場編）』『日本語日常会話コーパス』（2017 年 4 月時点での構築途中のもの）である。主な結果として，①品詞構成比では，名詞，副詞は，CSJ 学会講演と日常会話・名大とが対照的な分布を示し，また，終助詞，感動詞-フィラーは，CSJ 学会講演・CSJ 摸擬講演と日常会話・名大とが対照的な分布を示す（下の図 1）。②特徴語では，コーパス間で感動詞（一般，フィラー），終助詞，人称代名詞の分布に違いが見られた。また，これらの語の使用において，

性差の違いの方が年齢層の違いよりも特徴的な語数が多いことが観察された。

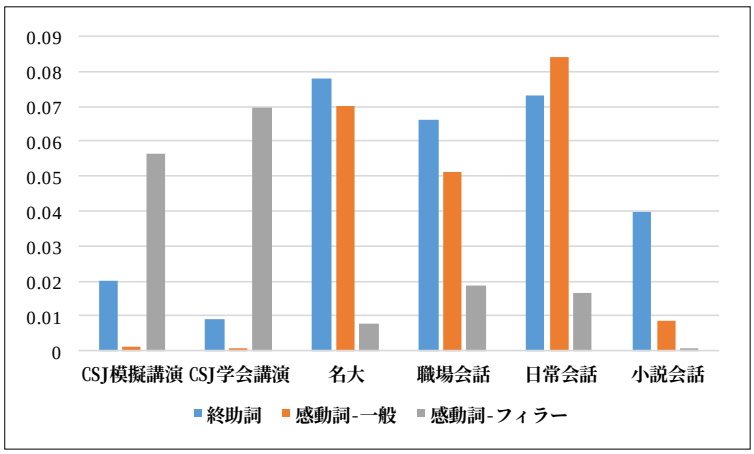

図１　終助詞，感動詞の割合

５．主な発表論文等

〔雑誌論文〕（計 7 件）

(1)山崎誠, 翻訳小説と日本語小説における会話文の計量語彙論的比較, 語彙研究, 15, pp.1-15, 2018，査読有
(2)Makoto Yamazaki, Yumi Miyazaki, Wakako Kashino, Annotation and Quantitative Analysis of Speaker Information in Novel Conversation Sentences in Japanese, 11th edition of the Language Resources and Evaluation Conference (LREC2018), 2018，査読有
https://www.aclweb.org/anthology/L18-1174
(3)財津亘，金明哲，性別を偽装した文章における文体的特徴変化，同志社大学ハリス理化学研究報告，59(3)，pp.47-54，2018，査読有
(4)財津亘，金明哲，文末語の使用率に基づいた筆者識別―探索的多変量解析の実施と分析結果に対するスコアリングによる検討―，計量国語学，31(6)，pp.417-425，2018，査読有
(5)財津亘，金明哲，テキストマイニングを用いた筆者識別へのスコアリング導入―文字数やテキスト数，文体的特徴が得点分布に及ぼす影響―，日本法科学技術学会誌，22(2)，pp. 91-108，2017，査読有
DOI:https://doi.org/10.3408/jafst.715
(6)山崎誠，コーパスが変える日本語の科学―日本語研究はどのように変わるか―，日本語学，35(12)，pp.12-17，2016，査読無
(7)山崎誠，基本統計量に現れるテキストの特徴，日本語学，34(7)，pp.78-83，2015，査読無

〔学会発表〕（計 17 件）

(1)入江さやか，金明哲，コーパスを用いた仮定形における音韻融合使用と印象評定に関する研究，シンポジウム「日常会話コーパス」Ⅳ，2019
(2)清水まさ子，日常会話と疑似会話における「って」の使用比較―引用・伝聞用法を中心に―，シンポジウム「日常会話コーパス」Ⅳ，2019
(3)柏野和佳子，日常会話の自称詞と小説会話の自称詞，シンポジウム「日常会話コーパス」Ⅳ，2019
(4)山崎誠，宮嵜由美，柏野和佳子，BCCWJ 小説会話文への発話者情報の付与と計量的分析，計量国語学会第 62 回大会，2018
(5)山崎誠，話し言葉における代名詞「あれ」の用法の分布，言語資源活用ワークショップ 2018，2018
(6)山崎誠，小説会話文への話者情報付与とその問題点，テキストマイニング 2018，2018
(7)山崎誠，翻訳小説における会話文の語彙的特徴，シンポジウム「日常会話」Ⅲ，2018
(8)柏野和佳子，フォーマルな話し言葉に現れやすい書き言葉的な語，シンポジウム「日常会話」Ⅲ，2018
(9)宮嵜由美，柏野和佳子，山崎誠，『現代日本語書き言葉均衡コーパス』収録の小説における発話箇所認定について，シンポジウム「日常会話」Ⅲ，2018
(10)李広微，金明哲，現代日本語小説の文体的特徴の変化について―大正・昭和の作品を中心として―，第 46 回日本行動計量学会，2018
(11)尾城奈緒子，金明哲，文末表現に着目した文学作品の分類，2018 年度日本分類学会シンポジウム，2018
(12)山崎誠，外国語翻訳小説と日本語小説の会話文の計量語彙論的比較，2017 年語彙研究会大会，2017
(13)黄善玉，金明哲，文型を特徴量とした文章の著者識別，第 45 回日本行動計量学会，2017
(14)山崎誠，レジスター・位相の違いによる会話文の語彙的多様性，言語資源活用ワークショ

ップ 2017，2017
DOI https://doi.org/10.15084/00001529
(15)宮嵜由美，柏野和佳子，山崎誠，発話文への発話者情報付与の設計―BCCWJ 収録の小説を対象に―，言語資源活用ワークショップ 2016，2017
DOI:info:doi/10.15084/00001456
(16)高崎みどり，文章・談話資料中の広義"疑似会話文"について，シンポジウム「日常会話コーパス」II
(17)宮嵜由美，山崎誠，柏野和佳子，『現代日本語書き言葉均衡コーパス』収録の小説を対象とした話者属性情報付与の検討，シンポジウム「日常会話コーパス」II

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：茂木俊伸
ローマ字氏名：(MOGI, Toshinobu)
所属研究機関名：熊本大学
部局名：大学院人文社会科学研究部（文）
職名：准教授
研究者番号（8 桁）：20392540

研究分担者氏名：柏野和佳子
ローマ字氏名：(KASHINO, Wakako)
所属研究機関名：大学共同利用機関法人人間文化研究機構国立国語研究所
部局名：音声言語研究領域
職名：准教授
研究者番号（8 桁）：50311147

研究分担者氏名：高崎みどり
ローマ字氏名：(TAKASAKI, Midori)
所属研究機関名：お茶の水女子大学

部局名：
職名：名誉教授
研究者番号（8 桁）：60096237

研究分担者氏名：金明哲
ローマ字氏名：(KIN, Meitetsu)
所属研究機関名：同志社大学
部局名：文化情報学部
職名：教授
研究者番号（8 桁）：60275469

研究分担者氏名：清水まさ子
ローマ字氏名：(SHIMIZU, Masako)
所属研究機関名：日本女子大学
部局名：文学部
職名：研究員
研究者番号（8 桁）：80649468

(2)研究協力者
研究協力者氏名：
ローマ字氏名：